La Macif, c'est vous.

Manifeste Macif pour l'utilisation éthique de l'Intelligence Artificielle

Elaboré par la Commission mixte gouvernance du numérique Macif

Janvier 2025

SOMMAIRE

Pourquoi un Manifeste Macif pour l’utilisation éthique de l’intelligence artificielle ?

Dans le contexte de la récente adoption de la réglementation européenne de l'intelligence artificielle, la Macif a souhaité répondre au besoin de positionnement des acteurs, en élaborant son *Manifeste Macif pour l'utilisation éthique des systèmes d'Intelligence Artificielle*.

Ce document expose :

1. Les raisons pour lesquelles le développement et l'utilisation de systèmes d'intelligence artificielle s'accompagnent d'un besoin de questionnement éthique,
2. Les enjeux éthiques relatifs à la mobilisation de systèmes d'intelligence artificielle dans le secteur de l'assurance,
3. La manière dont la Macif structure son questionnement éthique vis-à-vis de ces systèmes.

Par définition, ce document a vocation à évoluer dans le temps en fonction des avancées technologiques et des cadres normatifs qui se mettront progressivement en place.

Pourquoi un Manifeste Macif pour l'utilisation éthique de l'intelligence artificielle ?

Les systèmes d'intelligence artificielle qui se développent à l'heure actuelle constituent un défi majeur en termes d'évaluation et de maîtrise pour un assureur mutualiste comme la Macif.

La mutuelle ne veut pas se priver de l'apport potentiel des technologies d'IA pour améliorer sa performance relationnelle, opérationnelle, économique et sociale.

Depuis sa création, la Macif a toujours su renouveler ses approches, innover, proposer des solutions inédites, utiliser les outils les plus performants pour servir ses sociétaires, adhérents et clients mais aussi ses collaborateurs.

Elle entend poursuivre cette stratégie à l'ère de l'intelligence artificielle.

Elle n'ignore pas non plus que cette technologie, par sa puissance et la mauvaise utilisation qui peut en être faite, recèle d'importants dangers pour les valeurs qu'elle défend de manière affirmée et constante : inclusion, solidarité et lutte contre les discriminations notamment.

Pour cette raison, la Macif a souhaité adopter une approche novatrice, volontariste et réfléchie sur son utilisation future de l'intelligence artificielle.

Elle a souhaité que la réflexion précède l'action.
Elle a souhaité qu'un cadre politique soit mis en place pour faire émerger et croître les futurs usages qu'elle fera de l'IA.

Si elle a voulu que ce cadre soit solide sur le plan des valeurs et de l'éthique, elle a aussi souhaité qu'il soit évolutif et qu'il garde sa pertinence dans le temps : la Macif n'ignore pas que l'intelligence artificielle n'en est qu'à son émergence et que nous ne mesurons pas encore à quel point elle va transformer notre société, nos façons de vivre, nos façons de faire et de penser.

La mutuelle a donc décidé de se doter d'un texte fondateur, de poser un cadre initial pour la mise en œuvre de ses premiers cas d'usage de systèmes d'intelligence artificielle :

Le Manifeste Macif pour l'utilisation éthique de l'intelligence artificielle

Ce manifeste définit précisément ce qu'est l'intelligence artificielle, les risques majeurs qu'elle présente et les problématiques spécifiques qu'elle fait émerger ; mais il présente aussi l'apport énorme qu'elle peut avoir dans l'ensemble de la chaîne de valeur assurantielle.

Il met en lumière le fait qu'un acteur mutualiste comme la Macif ne peut pas traiter cette technologie comme un outil ordinaire car il met potentiellement en tension son cadre de valeur.

Le manifeste fixe la manière dont les équipes techniques doivent aborder les futurs projets impliquant l'intelligence artificielle et comment la mutuelle met en place un cadre de gouvernance pour s'assurer que la croissance du recours à cette technologie ne s'accompagne pas d'une croissance du risque pour elle, ses collaborateurs et ses sociétaires, adhérents et clients.

Le manifeste pose le principe d'une approche pragmatique et attentive du sujet, qui se caractérise notamment par :

- Une responsabilisation forte des techniciens qui seront en charge de ce sujet,
- Un questionnement permanent sur le bien-fondé d'utiliser l'IA à la place d'autres solutions dans les cas d'usage envisagés,
- Un regard multiple et complémentaire, à la fois technique et politique, sur les cas d'usage,
- Un principe de jurisprudence pour ne pas freiner la diffusion de cas d'usage au sein de la mutuelle,
- Un pilotage permanent et exhaustif de tous les projets et mises en œuvre de l'IA, quel que soit leur stade d'avancement, pour que l'entreprise et ses dirigeants soient en permanence informés de son utilisation au sein de la mutuelle.

Le manifeste décrit enfin les trois parties prenantes directes du sujet au sein de la Macif et leur rôle et responsabilité : le groupe de travail spécialisé, le GT-IA, la Commission mixte gouvernance du numérique et le Comité de Direction de la Macif. Ce dernier est évidemment, comme pour tous les autres sujets, responsable de ses décisions devant le Conseil d'administration de la Macif.

Avec ce manifeste, qui sera régulièrement questionné et évoluera dans le futur, et la gouvernance qu'elle met en place sur le sujet, la Macif peut sereinement s'engager sur le chemin de l'utilisation des systèmes d'intelligence artificielle et en faire un nouveau moteur de performance au service de ses parties prenantes et des valeurs qu'elle porte.

1. Les enjeux éthiques de l'intelligence artificielle

1.1 Les systèmes d'intelligence artificielle et leurs technologies

La notion d'intelligence artificielle (IA) est floue[1]. Il n'existe pas et n'a jamais existé de définition consensuelle de ce qu'est précisément l'IA, notamment car la notion d'intelligence elle-même est soumise à débat : selon leurs disciplines, leurs métiers et leurs objectifs, les chercheurs, les développeurs et les praticiens lui donnent, chacun, un sens différent.

Dans ce manifeste, **l'intelligence artificielle est définie comme la capacité, pour une machine, de produire des tâches cognitives similaires à celles des humains** comme l'apprentissage, le raisonnement, la résolution de problèmes ou encore le traitement du langage.
Ces tâches incluent la description, la modélisation et l'évaluation de données, l'élaboration de prédictions, la production de recommandations, la prise de décision, la planification et le traitement d'opérations, la production de contenu de multiple nature - textuelle, sonore, visuelle, graphique, audiovisuelle...

A proprement parler, il n'existe pas à ce jour d'intelligence artificielle au sens où une machine serait capable de se comporter comme le ferait un humain dans ses différents contextes sociaux et de déployer l'ensemble de ses tâches cognitives de manière réflexive, située et créative. Il existe seulement des systèmes d'IA qui se définissent, à la suite de l'OCDE et du règlement européen sur l'intelligence artificielle, l'*AI Act*, comme des programmes destinés à répondre de manière plus ou moins autonome à des objectifs - précis ou généraux - fixés par l'humain, qui mettent en oeuvre les capacités nécessaires pour atteindre ces objectifs et qui peuvent, éventuellement, adapter leur fonctionnement après leur déploiement[2].

[1] Dans le cadre de la construction de la réglementation européenne AI Act, la question de la définition de l'IA a précisément fait l'objet de longues discussions.

[2] Dans l'AI Act, les systèmes d'IA sont définis comme des systèmes automatisés conçus pour fonctionner à différents niveaux d'autonomie et pouvant présenter une capacité d'adaptation après leur déploiement et qui, pour des objectifs explicites ou implicites, déduisent de la contribution reçue la manière de générer des résultats tels que des contenus, des prédictions, des recommandations ou des décisions qui peuvent influencer les environnements physiques ou virtuels.

Les systèmes d'intelligence artificielle se distinguent des systèmes informatiques traditionnels. Dans le cas des systèmes classiques, les règles du traitement de l'information sont déclarées par les êtres humains qui les développent. Dans le cas des systèmes d'IA, ces règles sont inférées à partir des données. Le développement puis l'utilisation des systèmes d'IA s'appuient sur l'analyse de grands ensembles de données provenant, selon les besoins, de sources diverses : données de marchés, d'activités, de consommation, de navigation internet, données issues des réseaux sociaux ou encore d'objets connectés, etc. Les systèmes d'IA sont construits sur la base d'un apprentissage automatique, par les machines, des traitements qu'elles doivent opérer sur les données traitées.

L'apprentissage automatique désigne la capacité, pour une machine, d'apprendre à partir de grands ensembles de données. Il consiste à mobiliser diverses méthodes statistiques et d'optimisation dans le but d'identifier des régularités dans les données. Cet apprentissage peut se dérouler de plusieurs façons[3]:

- **Il peut tout d'abord être supervisé par l'humain** : il s'agit pour la machine d'apprendre à prédire des résultats sur la base de ce que l'humain désigne comme étant les bons résultats dans un ensemble de données appelé échantillon. Une fois l'algorithme suffisamment robuste, celui-ci est utilisé pour appliquer sa fonction sur de nouvelles données.

- **L'apprentissage automatique peut ensuite être non supervisé.** Dans ce cas, les données ne sont pas étiquetées par l'humain et la machine tâche d'identifier les caractéristiques communes aux données pour les classer et en modéliser les structures sous-jacentes. Il est par ailleurs possible de combiner les deux grandes approches supervisées et non supervisées dans le cadre d'approches dites partiellement supervisées, semi-supervisées ou auto-supervisées.

- Les capacités d'un premier modèle peuvent également être utilisées dans le cadre d'un second modèle nécessitant des compétences similaires à celles déjà apprises : **c'est l'apprentissage par transfert.**

- **Enfin, l'apprentissage automatique peut se dérouler par renforcement** : il s'agit d'un apprentissage comportemental proche de l'apprentissage supervisé dans lequel un agent logiciel apprend de ses expériences (essais et erreurs) en fonction des réponses positives et négatives qu'il reçoit de la part d'un humain ou bien d'un autre algorithme.

[3] Par souci de simplicité, seules les approches les plus communes sont ici citées.

Sur la base de ces mécanismes d'apprentissage, plusieurs technologies d'intelligence artificielle ont été développées ces dernières années. Les technologies d'apprentissage dit profond - une forme d'apprentissage automatique qui mobilise des technologies de neurones artificiels multi-couche possédant un très grand nombre de paramètres[4] - ont permis le développement de systèmes de traitement et de génération de langage naturel et d'autres formes de contenus qui sont aujourd'hui largement diffusés :

- **Les grands modèles de langage - ou "LLM" -**, technologies de traitement du langage naturel par analyse lexicale et statistique des textes, sont utilisés dans les agents conversationnels de dernière génération[5]. Ces derniers sont alors capables de proposer une réponse à toute question posée en l'extrapolant à l'aide de méthodes statistiques, **sans garantir que cette réponse soit nécessairement juste** : le système prédit le prochain caractère le plus probable de la réponse sur la base d'un historique non supervisé et son comportement se trouve renforcé par une procédure d'alignement en fonction des retours d'expérience des humains, de manière à ce que la réponse produite soit la plus conforme à ce qui est attendu ;
- **Les réseaux antagonistes génératifs - ou « GAN » -** technologies d'apprentissage où un réseau de neurones artificiels est entraîné à produire du contenu face à un second réseau capable de reconnaître le caractère vraisemblable ou non de ce contenu, ont pour leur part permis le développement de modèles de génération d'images, de vidéo et d'audio[6].

Ces technologies les plus récentes ont considérablement renforcé la puissance des systèmes d'IA et ont provoqué un saut qualitatif sans précédent. Celui-ci s'est accompagné d'une prise de conscience, par le grand public, des capacités offertes par l'intelligence artificielle - notamment sur le volet génératif - et de la nécessité de se questionner sur ses opportunités et ses dangers.

4 Un modèle d'origine chinoise publié en mars 2022 a été entraîné sur 174 milliers de milliards de paramètres.

5 Le modèle GPT-3 utilisé lors de son lancement par ChatGPT, qui a popularisé ce type d'agents conversationnels, a été entraîné sur 175 milliards de paramètres.

6 Technologies qui sont, notamment, à la base des « deepfakes », contrefaçons profondes.

1.2 Mobiliser des systèmes d'IA de manière responsable : l'enjeu de l'équité des systèmes

Les algorithmes peuvent renforcer des biais, conscients ou non, qui conduisent à désavantager ou exclure une partie de la population dans le cadre des tâches qu'ils mettent en œuvre. L'existence de telles situations non souhaitables dans les algorithmes a été plusieurs fois observée durant la dernière décennie (par exemple s'agissant d'algorithmes de tarification ou d'algorithmes de justice prédictive), si bien que l'association professionnelle en charge de la production des normes applicables dans les secteurs de l'informatique et des télécommunications a engagé, dès 2016, un travail mondial sur la conception d'une intelligence artificielle équitable.

L'équité algorithmique se définit comme « l'absence de préjudice ou de favoritisme envers un individu ou un groupe en raison de ses caractéristiques innées ou acquises ». Les préjudices et les favoritismes sont la cause de biais qui peuvent être introduits de trois manières dans les algorithmes.

1. Des biais de comportements humains peuvent se trouver dans les données d'apprentissage et être incorporés dans les algorithmes. Pour une intelligence artificielle, les données qui constituent son univers d'apprentissage sont des faits bruts qu'elle n'évalue pas moralement. A partir de ces données, l'IA modélise un comportement qui reproduit mécaniquement les représentations, les critères de jugement et les préférences humaines qui s'y trouvent. Lorsque ces derniers sont biaisés, les biais en question sont internalisés dans les systèmes qui les mettent ensuite en œuvre dans le traitement des tâches pour lesquelles ils ont été conçus. **Ce mécanisme s'appelle le biais historique**.

2. Des biais méthodologiques dans la phase de constitution de l'ensemble des données d'apprentissage peuvent engendrer des anomalies dans les algorithmes. Ceux-ci peuvent par exemple consister, pour une intelligence artificielle, à s'entraîner sur la base d'un échantillon de données qui n'est pas représentatif de l'ensemble des données qu'elle aura à traiter dans le cadre de sa fonction (biais de sélection), à omettre une ou plusieurs variables explicatives des phénomènes à modéliser ou à utiliser de mauvaises variables proxy pour les mesurer, à tirer des conclusions au niveau individuel de résultats agrégés au niveau d'une population (erreur écologique), etc. Les biais méthodologiques sont bien connus des *data scientists* qui doivent contrôler leur absence durant le processus de conception des systèmes d'IA.

3. Des biais de comportement des utilisateurs peuvent être provoqués par l'usage des systèmes d'IA. La manière dont les résultats d'un algorithme sont triés puis présentés par l'interface homme-machine influence grandement l'attention des utilisateurs, leurs choix, leurs parcours et leurs actes (biais d'ancrage, biais du survivant, biais de désirabilité sociale, mimétisme, etc.). En retour, le comportement des utilisateurs peut renforcer l'algorithme dans son fonctionnement et ainsi de suite. De manière générale, l'apprentissage par renforcement rend possible l'émergence de biais qui n'existaient pas avant la mise en production de l'algorithme. Ces biais sont le plus souvent l'effet d'une évolution dans la population des utilisateurs entre la phase d'apprentissage et la livraison de l'outil incorporant l'IA. Il arrive également qu'ils soient délibérément introduits par

des utilisateurs malveillants (dans le cadre d'une attaque dite adverse).

De nombreuses mesures techniques peuvent être mises en œuvre dans l'ensemble de la phase de conception et d'élaboration des systèmes afin de détecter et prévenir le risque de discrimination algorithmique. Celles-ci peuvent porter sur les données d'apprentissage (suppression des biais existants), sur le processus d'apprentissage (imposition de contraintes particulières) ou encore sur l'algorithme qui en est issu (modification de certaines fonctions). Cependant, ces mesures ne règlent pas l'ensemble du problème car **le problème de l'équité algorithmique n'est pas seulement d'ordre technique**.

En effet, la discrimination algorithmique n'est pas uniquement le résultat de biais objectifs à l'image des biais méthodologiques. Elle est en majeure partie construite et située, c'est-à-dire relative aux modèles de justice existants dans des contextes socio-historiques donnés : ce que l'on reconnaît comme équitable ou inéquitable dépend du point de vue adopté. Que considère-t-on comme équitable ? A quel moment juge-t-on qu'il y a discrimination ou exclusion ? Les modèles de justice sont pluriels et constituent autant de schémas de raisonnement qui aboutissent à des prises de décision divergentes voire contradictoires entre elles.

Par ailleurs, ce qui est considéré comme acceptable ou discriminatoire évolue dans le temps[7] et diffère selon les populations et les cultures.

Les systèmes d'IA ne peuvent pas satisfaire à tous les critères d'équité et de solidarité en même temps.

Un algorithme n'est jamais « juste » car « la » justice n'existe pas en tant que telle ; il existe seulement des conceptions particulières de ce qui est juste, équitable, responsable, solidaire, en fonction des cas d'usages et des publics.

Il convient donc, pour les entreprises qui développent et utilisent des systèmes d'IA au sein de leurs chaînes de valeur, d'arbitrer entre les différentes possibilités de traitement algorithmique en mobilisant les théories de la justice jugées appropriées pour servir leurs intentions politiques et stratégiques puis d'édicter les critères extra-techniques à prendre en compte dans le design des systèmes d'algorithmes et d'intelligence artificielle.

Si identifier et caractériser les biais méthodologiques est un exercice technique, définir ce qui relève de la discrimination, de l'inéquité, de l'injustice ou de l'exclusion est un exercice éthique et politique.

Il s'agit d'un exercice éthique car les entreprises qui mobilisent des systèmes d'IA endossent la responsabilité des traitements qu'elles appliquent aux personnes qui sont l'objet des systèmes, en particulier leurs clients et collaborateurs. Il s'agit d'un exercice politique car les systèmes d'IA contribuent à construire le futur de l'activité de ces entreprises en faisant évoluer leurs métiers, leurs offres, ainsi que les comportements de consommation.

[7] Par exemple, le 2 mars 2011, la Cour de justice de l'Union européenne a rendu un arrêt mettant fin à la différence de tarification appliquée aux jeunes conducteurs et aux jeunes conductrices. Cette inégalité, qui consistait à faire payer plus cher les hommes que les femmes en raison de la plus grande sinistralité du premier groupe, a été jugée discriminatoire.

1.3 Comprendre pour décider et rendre compte : l'enjeu de l'explicabilité des systèmes

Contrairement aux systèmes classiques où les humains déclarent explicitement les critères successifs du traitement de l'information, les systèmes d'IA fonctionnent sur la base de règles inférées des données.

Dans les systèmes d'IA modernes conçus sur la base de technologies d'apprentissage profond, ces règles sont difficilement, voire pas du tout, interprétables par les humains.

Les systèmes d'IA sont capables de produire les résultats attendus sans que l'on sache comment ils y parviennent : les systèmes sont dits opaques car les traitements qu'ils opèrent sur les données sont des boîtes noires. Or, les entreprises qui mobilisent des systèmes d'IA peuvent vouloir s'assurer que les décisions et traitements opérés de manière automatique sont conformes à leurs valeurs et à leurs intentions stratégiques ; elles doivent pour cela rendre compte de ces opérations, particulièrement lorsque les systèmes d'IA sont utilisés dans des composantes critiques de leurs chaînes de valeur.

L'enjeu de la compréhension des systèmes par les humains est triple. Si celle-ci n'est pas forcément obligatoire selon les cas d'usage, elle est souhaitable dès lors qu'il s'agit de gagner l'adhésion des parties prenantes au système, d'anticiper les comportements du système dans les situations marginales et d'être capable, au niveau de chacune des prédictions individuelles rendues par le système, d'expliquer comment la prédiction en question est produite (pour exercer ou se défendre de recours de tiers, par exemple).

Pour répondre au problème de l'opacité des systèmes d'IA, deux grandes voies peuvent être envisagées en fonction des niveaux de compréhension et de responsabilité recherchés.

La première voie consiste à employer des technologies d'intelligence artificielle interprétables. Ces technologies spécifiques, quoiqu'encore généralement moins performantes que les technologies d'apprentissage profond, ont été développées ces dernières années afin de permettre d'aboutir à des systèmes transparents dont les humains peuvent comprendre le fonctionnement : les systèmes intrinsèquement interprétables sont construits avec un nombre réduit de paramètres qui font sens pour leurs utilisateurs. L'interprétabilité des systèmes peut être globale et offrir une compréhension de l'entièreté du système ou bien locale et offrir une compréhension des fonctions particulières du système.

Dans les cas où des systèmes d'IA sont utilisés dans des composantes critiques de l'activité d'une entreprise, celle-ci peut faire le choix, quand cela est faisable, de concevoir ces systèmes de manière à ce qu'ils soient intrinsèquement interprétables.

La seconde voie consiste à construire *a posteriori* une explication suffisamment satisfaisante du fonctionnement d'un système opaque.

Les systèmes conçus sur la base de technologies d'apprentissage profond mobilisent un nombre de paramètres si élevé (jusqu'à des centaines de milliards de paramètres pour les grands modèles

de langage) qu'il est impossible d'en comprendre le fonctionnement interne. En revanche, il est possible de fournir certaines explications aux résultats que le système produit. Ces explications données *a posteriori* peuvent par exemple consister à fournir une modélisation simplifiée - une approximation – du fonctionnement du système au sujet d'un résultat particulier qu'il produit. L'explication qu'il s'agit de construire dépend du niveau de compréhension visé, ainsi que du type de public auquel il s'agit de rendre des comptes. **Expliquer signifie à la fois donner une réponse au « comment » un système d'IA fonctionne, ainsi qu'au « pourquoi » un système donne tel ou tel résultat.**

La question du « comment » concerne l'auditabilité technique du système : celui-ci respecte-t-il le cahier des charges, est-il efficace dans les processus métiers qu'il sert et est-il conforme à la réglementation ? La question du « pourquoi » revêt, quant à elle, des enjeux d'ordre éthique. Il s'agit, pour les opérateurs humains qui utilisent le système, d'en comprendre le comportement afin de l'utiliser à bon escient, de s'assurer que celui-ci est en phase avec les valeurs de l'entreprise et que celui-ci ne comporte pas de biais discriminatoires. Il s'agit, pour les clients qui en sont les sujets, de comprendre pour quels motifs le système rend telle recommandation ou telle décision à leur égard.

L'explicabilité des systèmes d'IA n'est pas seulement nécessaire pour s'assurer que ceux-ci sont en accord avec les types de justice, d'équité, d'inclusion et de solidarité voulus par l'entreprise qui les mobilise dans son activité. **Cette explicabilité est aussi nécessaire pour maintenir dans le temps un contrôle de ces systèmes par les humains.**
S'assurer d'une capacité de contrôle effective des systèmes d'IA implique que les opérateurs humains qui les utilisent disposent des capacités cognitives et attentionnelles suffisantes pour connaître le fonctionnement et les limites de ces systèmes, éventuellement remettre en question les résultats produits par l'IA et, au besoin, reprendre la main sur le processus. Les types d'interactions homme-machine et les types de contrôles des systèmes par l'humain qu'il s'agit de mettre en œuvre dépendent là encore des cas d'usage et de la criticité des systèmes en question dans les chaînes de valeur. **Enfin, la responsabilité des entreprises vis-à-vis des systèmes qu'elles mobilisent nécessite qu'elles traitent de manière différenciée ces systèmes selon leurs origines ainsi que selon les technologies qu'elles incorporent.**

Selon que ces systèmes sont développés en interne, des produits génériques adaptés aux usages internes ou encore des produits externes "sur étagère", leur degré d'explicabilité et le niveau de contrôle qu'il faut leur appliquer peuvent varier. Il en est de même selon les technologies utilisées pour construire ces systèmes. **Il est ainsi nécessaire de porter une attention toute particulière aux modèles fondationnels à usage général** (tels que ceux utilisés par les IA génératives) dont le fonctionnement est fondamentalement opaque. Leurs données d'entraînement sont inconnues et recèlent de nombreux biais. Les systèmes créés sont capables d'hallucinations et leurs comportements ne peuvent pas être contraints à l'aide de règles explicitement déclarées par les entreprises qui les déploient.

Ainsi, il s'agit pour les entreprises de mobiliser les systèmes d'IA en connaissance de cause selon le niveau de responsabilité qu'elles entendent endosser et donc de la criticité des sages visés.

2. L'intelligence artificielle, une nouvelle donne dans l'industrie assurantielle

2.1 La transformation des métiers de l'assurance sous l'effet de l'intelligence artificielle

La maturité des technologies d'IA est aujourd'hui telle que celles-ci ont désormais le potentiel de faire évoluer en profondeur les économies et les sociétés en repoussant très significativement le domaine du possible s'agissant des capacités de modélisation de tous types de données et d'automatisation des tâches. Ces dernières années, certaines industries ont vu se diffuser largement des systèmes d'IA dans l'ensemble des composantes de leurs chaînes de valeur. C'est notamment le cas de la banque et de la finance, de l'automobile, du secteur médical, de la production manufacturière et de la logistique ou encore de l'économie du partage. **Le secteur assurantiel, pour sa part, débute sa mue**.

Les applications - déjà opérationnelles ou futures- de l'IA dans l'assurance déploient trois ensembles de capacités :

- Des capacités de reconnaissance, de traitement et de génération de langage naturel (écrit, parlé, corporel), utilisées par exemple par les robots conversationnels jouant le rôle d'assistants virtuels au service des clients et des conseillers ;
- Des capacités de modélisation de données et de prédiction de tendances (comme les comportements et les préférences), utilisées par exemple pour améliorer l'efficacité opérationnelle en anticipant les volumes d'activité ;
- Des capacités de traitement de contenu (audio, visuel, biométrique, financier) et de prise de décision

automatisée, utilisées par exemple pour l'identification, l'évaluation, la gestion, la prévention et le règlement des risques et sinistres pesant sur ou ayant affecté des personnes, des objets, des actifs ou des territoires.

Potentiellement, ces trois ensembles de capacités sont mobilisables dans l'ensemble des composantes de la chaîne de valeur de l'assurance, que celle-ci concerne les biens ou les personnes. Ainsi, l'ensemble des activités que les assureurs conduisent, activités métiers et activités support, peut bénéficier des nouvelles capacités offertes par les systèmes d'IA.

S'agissant des activités de développement et de mise sur le marché de nouveaux produits (étude de marché, analyse des segments de clientèle, tarification des produits et communication), les capacités d'analyse prédictive et de détection des tendances, couplées à l'utilisation de moteurs de recommandation, peuvent permettre aux assureurs de personnaliser la manière dont ils s'adressent à leurs clients et de personnaliser leurs offres, tant dans leurs niveaux de garanties que dans leurs prix.
S'agissant du conseil, de la vente et de la gestion de la relation client, les capacités de traitement du langage naturel alliées aux capacités d'analyse précédemment citées peuvent permettre d'assister les conseillers en fournissant des recommandations sur les besoins de leurs assurés, voire d'automatiser les activités de conseil et vente tout en améliorant la qualité de la relation.
S'agissant de l'administration des contrats et des services rendus aux assurés, les capacités déjà évoquées, ainsi que celles plus spécifiquement liées à la prévention des risques, peuvent permettre de rendre plus efficients, voire d'automatiser, les processus de gestion des demandes, des sinistres et des réclamations, de lutter plus efficacement contre la fraude ou encore de fournir des éléments de connaissance personnalisés quant aux risques encourus par les assurés.

Les assureurs pourraient aussi utiliser l'IA pour hybrider de plus en plus fortement l'approche historique de réparation des sinistres par une démarche technologiquement renouvelée de prédiction et de prévention des risques.
En somme, les systèmes d'IA pourraient être très largement utilisés par les assureurs pour créer une meilleure expérience client, innover dans leurs produits et leurs services, gérer leurs flux de manière plus efficiente, réduire leurs coûts, réduire leurs risques et augmenter leur développement.

2.2 Une vigilance éthique nécessaire face au développement de l'IA dans l'assurance

Si la mobilisation de systèmes d'IA dans l'activité des assureurs apparaît aujourd'hui clairement comme une opportunité d'améliorer le service et la protection des assurés, celle-ci doit être envisagée avec vigilance. En effet, plusieurs dérives possibles peuvent être identifiées dans chacune des composantes de la chaîne de valeur et des questions éthiques doivent dès lors être soulevées.

S'agissant des activités de développement et de mise sur le marché de nouveaux produits, une micro-segmentation rendue possible par l'analyse des données personnelles peut éventuellement aboutir à une remise en cause des mécanismes classiques de solidarité entre les assurés.

Elle peut conduire, sur le marché, à l'exclusion des personnes les plus exposées ou les plus fragiles du fait d'une tarification prohibitive des produits d'assurance sur certains micro-segments.

S'agissant du conseil, de la vente et de la gestion de la relation client, une trop grande atomisation des offres pourrait s'accompagner d'une plus grande difficulté à équiper correctement les assurés, d'une sélection biaisée des nouveaux clients sur les micro-segments les plus profitables ou encore de la généralisation de mécanismes de *lock-in* empêchant les assurés de faire valablement jouer la concurrence.

S'agissant de l'administration des contrats et des services rendus aux assurés, une automatisation des processus pourrait conduire à discriminer les personnes dont les compétences technologiques sont moindres, à effectuer des modifications de contrats ou des ventes inadaptées ou à rendre des décisions biaisées au détriment de certains groupes de populations spécifiques. Cela est par exemple le cas d'une détection algorithmique des fraudes fondée sur un profilage des cas antérieurs avérés.

En somme, les technologies d'apprentissage automatique s'accompagnent de nouveaux risques et donc de nouveaux enjeux : des enjeux éthiques relatifs à l'équité de traitement des personnes, à la non-discrimination et l'inclusion des publics, ainsi que des enjeux techniques relatifs à l'explicabilité des algorithmes, à leur bonne compréhension et à leur mobilisation responsable par les assureurs qui les utilisent. Ils appellent de leur part la mise en œuvre de travaux réflexifs, prospectifs et normatifs sur l'usage que les assureurs entendent faire de ces technologies et de se doter des mécanismes de gouvernance appropriés pour éviter tout risque de dérive ou de mésusage.

2.3 Une vigilance éthique supplémentaire pour les assureurs mutualistes

Pour les assureurs mutualistes, les questions éthiques qui accompagnent la conception et le déploiement de systèmes d'IA dans l'industrie assurantielle doivent faire l'objet d'une vigilance supplémentaire.

Le mutualisme trouve son origine dans la volonté de personnes de construire une réponse collective à des besoins de protection individuels. Sa raison d'être est de prendre en charge ces besoins en conduisant une activité assurantielle au seul bénéfice des sociétaires[8], c'est-à-dire en faisant en sorte de ne jamais capter ou détruire de la valeur à leur détriment, par opposition aux assureurs à but lucratif. Le choix d'une forme mutualiste d'entreprise, c'est à-dire non-lucrative et à gouvernance démocratique, exprime une double conviction : une logique de maximisation du profit n'a pas sa place dans la protection des personnes (ainsi que celle de leurs biens) et la solidarité construite entre les sociétaires est une affaire de sort commun. La vision politique portée par le mutualisme est ainsi celle d'une assurance la plus accessible et inclusive possible, la protection de tous servant la liberté de chacun au sein de la société. **Cette vision politique de l'assurance trouve à la fois des risques et des opportunités dans le développement des technologies d'intelligence artificielle.**

Lorsque les systèmes d'IA sont utilisés à des fins d'une meilleure identification et connaissance des risques, ceux-ci peuvent éventuellement être mis au service d'une intention prédatrice qui consisterait à exclure les mauvais risques pour ne garder que les bons. Les systèmes d'IA permettant de mieux distinguer les groupes et les personnes porteuses de risques accrus, certains assureurs pourraient mettre en place des stratégies de démutualisation visant à proposer des tarifs plus avantageux aux personnes les moins vulnérables. Le potentiel d'exclusion augmente donc avec la sophistication progressive des systèmes d'IA.

A l'inverse, une connaissance plus fine voire quasi individualisée des risques encourus par les sociétaires - connaissance permise par l'exploitation et le traitement de données issues de capteurs, de données de comportements de consommation, de données issues de l'*open data* public, etc. - peut être vue comme une opportunité lorsqu'elle est mise au service d'une meilleure prévention de ces risques, qu'ils relèvent de la santé des personnes ou des dommages à leurs biens. Les assureurs mutualistes peuvent donc y trouver une opportunité d'agir en amont de la réparation et de l'indemnisation et travailler à une meilleure protection et un meilleur accompagnement de leurs sociétaires.

8 On entend ici « sociétaire » dans le sens large de membre d'une société mutuelle, ce qui inclut les adhérents des mutuelles santé.

Pour un assureur mutualiste, les opportunités offertes par les systèmes d'IA sont à construire dans un cadre respectant l'autonomie des sociétaires. La surveillance des personnes à travers la collecte et l'analyse de leurs données personnelles doit demeurer raisonnable, c'est-à-dire réalisée dans le but d'apporter des gains significatifs en termes de prévention et de protection et elle doit recueillir le consentement éclairé des personnes qui en sont les sujets. A ce titre, la construction de mécanismes incitatifs de modification des comportements en utilisant l'IA doit demeurer inclusive et respectueuse de la liberté de choix des personnes. Accompagner les publics les plus exposés et les plus fragiles plutôt que de les exclure des mécanismes de protection collective est une exigence fondamentale de l'assureur mutualiste.

Enfin, pour un assureur mutualiste, les opportunités offertes par les systèmes d'IA ne peuvent pas être évaluées sans un examen de leurs **impacts environnementaux et sociaux**. L'entraînement et l'utilisation de systèmes d'IA sont très consommateurs de ressources énergétiques (cela est particulièrement le cas des IA génératives) : pour chaque utilisation envisagée de l'IA, il est donc nécessaire d'évaluer les bénéfices marginaux retirés par le recours à cette technologie, plutôt qu'aux systèmes classiques, au regard de leurs impacts environnementaux. Il est également nécessaire d'évaluer les systèmes d'IA sous l'angle de la transformation du travail des collaborateurs, c'est-à-dire questionner la manière dont ceux-ci s'insèrent dans le contenu d'un poste, d'une tâche, la manière dont ceux-ci peuvent éventuellement faire évoluer les identités professionnelles ou recomposer les interactions et relations humaines existantes.

3. L'éthique des systèmes d'intelligence artificielle à la Macif

3.1 L'approche Macif de l'éthique appliquée aux systèmes d'IA

La gouvernance du numérique au sein de la société évolue à l'articulation de trois ensembles :

- **la régulation du numérique**, qui consiste pour la puissance publique à fixer dans la loi et les cadres réglementaires les grands principes auxquels doivent se conformer l'ensemble des acteurs ;
- **la gouvernance du numérique des entreprises**, qui consiste pour elles à établir et implémenter les règles, procédures et standards s'agissant de la qualité et de la sécurité de leurs données, ainsi que de la manière d'exploiter ces données dans leurs activités ;
- **l'éthique du numérique**, qui consiste pour les entreprises à identifier les problèmes moraux relatifs aux différentes voies d'exploitation possibles de leurs données et à fixer la conduite qui exprime le mieux leurs valeurs.

Ainsi, la gouvernance du numérique des entreprises ne consiste pas seulement pour elles à se mettre en conformité avec ce que dictent la législation et la réglementation. Si la puissance publique fixe les règles à un moment donné, il appartient aux entreprises de se questionner et d'évaluer au fil du temps les options technologiques possibles s'agissant de l'exploitation de leurs données et notamment de la mobilisation qu'elles peuvent faire des systèmes d'intelligence artificielle.

Quels choix technologiques adopter pour construire le futur que nous, Macif, estimons être le plus désirable ?

Pour instruire les sujets d'éthique des systèmes d'intelligence artificielle, la Macif met en œuvre une procédure d'explicitation des enjeux relatifs à leurs usages puis de délibération des choix de conception et d'utilisation possibles de ces systèmes, ainsi que de leurs impacts une fois ceux-ci déployés.

En ce sens, à la Macif, l'évaluation de l'utilisation éthique ou non de l'intelligence artificielle est toujours appliquée aux situations concrètes et à visée pratique : elle prend pour objet les cas d'usage où l'entreprise envisage d'utiliser des systèmes d'IA pour gagner en performance.

Cette analyse éthique mobilise de façon complémentaire trois types de raisonnement :

1. Le raisonnement conséquentialiste

Il s'agit de considérer, pour les évaluer, les effets connus ou anticipés de l'usage des systèmes en termes organisationnels, économiques, stratégiques, politiques et sociétaux :

- Quels avantages les systèmes d'IA considérés confèrent-ils sur le plan de la performance opérationnelle, de la qualité de service, de la qualité des offres et quels avantages concurrentiels confèrent-ils à la Macif ?
- Quelles sont les évolutions des métiers de l'assurance et des comportements de consommation qui sont introduites par l'utilisation des systèmes d'IA considérés ?
- Quels risques particuliers en termes d'acceptabilité sociale et sociétale (usages, empreintes, soutenabilité, etc.) les systèmes d'IA considérés portent-ils ?

2. Le raisonnement déontologique

Il s'agit de mobiliser dans l'évaluation des systèmes d'IA la réglementation et les normes de la profession (en constante évolution), ainsi que les critères d'évaluation propres à la Macif :

- Les systèmes d'IA doivent être conformes à l'esprit de la loi, respecter les normes de la profession et être en phase avec les standards et usages ;
- Les systèmes d'IA considérés doivent permettre de mieux servir les sociétaires de la Macif en général ou bien des publics en particulier ;
- Les systèmes d'IA considérés doivent permettre d'améliorer l'expérience collaborateur ou encore les relations avec les parties prenantes de la Macif.

3. Le raisonnement axiologique

Il s'agit d'interroger le sens et le devenir des concepts fondamentaux de l'activité assurantielle, ainsi que des valeurs portées par la Macif dans son industrie et plus largement dans la société :

- Les systèmes d'IA considérés portent-ils une évolution des pratiques de mutualisation et de sélection des risques, d'uniformisation et d'individualisation des réponses aux besoins ?
- L'automatisation des pratiques et des processus permise par l'utilisation des systèmes d'IA se fait-elle au détriment ou au bénéfice d'une relation plus humaine aux sociétaires ?
- Les systèmes d'IA considérés permettent-ils de mettre en œuvre une plus grande solidarité entre les sociétaires et une plus grande accessibilité des offres de protection et d'accompagnement ?

A travers l'ensemble de ces questions, la Macif analyse et évalue l'intérêt, la pertinence et le bien-fondé de faire évoluer ses activités et ses offres au regard des nouvelles capacités offertes par les technologies fondées sur l'IA.

Pour rendre opératoire ce questionnement, la Macif suit un processus d'instruction bien défini tout au long duquel elle fait en sorte :

- D'expliciter les éventuels risques pouvant peser sur l'ensemble des populations qui interagissent avec un traitement automatisé et plus généralement toute question d'ordre éthique que soulèverait un cas d'usage,

- De mettre en regard les risques identifiés avec les bénéfices attendus,

- De définir, le cas échéant, les populations jugées sensibles et pour lesquelles un effort de protection particulier doit être apporté durant la phase de conception des systèmes d'IA,

- De contrôler ces systèmes avant leur mise en production puis de suivre leur déploiement au cours du temps.

3.2 Le cadre de gouvernance des systèmes d'IA à la Macif

La Macif adopte un processus général de développement et de contrôle des systèmes d'IA en cinq phases :

- L'évaluation préalable des risques du système,
- Son design,
- Son développement et son test,
- Son déploiement,
- Son contrôle une fois le système déployé.

Ce processus n'est pas linéaire mais itératif. La définition des attentes dans chacune des phases, ainsi que les différents points de validation, peuvent évoluer à la faveur d'une évaluation continue de la criticité du système tout au long du processus. Ainsi, par exemple, une évaluation supplémentaire des risques d'un système peut être demandée à la suite de l'apparition de doutes durant la phase de test de ce système.

Sous le contrôle du Conseil d'administration, la gouvernance de ce processus implique directement trois acteurs :

- **Le GT-IA** est le groupe de travail opérationnel qui recueille les besoins des utilisateurs potentiels des systèmes d'IA dans l'entreprise, les analyse, en fait une approche critique, suit la conception, le développement et le contrôle des systèmes d'IA à la Macif,

- **La Commission Mixte Gouvernance du Numérique** est l'instance réunissant des techniciens et des administrateurs de la Macif ; elle peut être mobilisée pour instruire un questionnement éthique dès lors qu'un système en phase de conception ou déployé soulève des interrogations,

- **Le Comité de Direction de la Macif** fixe les ambitions et limites stratégiques de l'utilisation des systèmes d'IA dans l'entreprise, les priorités d'usage et elle prend la responsabilité de déployer ou décommissionner un système d'IA à la suite d'une instruction éthique.

Le GT-IA a la charge de constituer une cartographie exhaustive des initiatives et projets d'IA de la Macif. L'instruction des cas d'usage adresse trois dimensions d'analyse principales : l'impact sociétal, l'impact environnemental et l'impact social des systèmes d'IA.

S'agissant de l'impact sociétal des systèmes d'IA, l'équipe en charge du design tâche d'expliciter et de limiter les risques de discrimination algorithmique.

A cette fin et pour chaque cas d'usage, l'équipe en charge du design du système d'IA s'attache à identifier et documenter sous un format standardisé et défini par le GT-IA :

- L'objectif du système
- Les utilisateurs cibles du système
- Les parties prenantes du système
- Les gains attendus et effets néfastes potentiels à minimiser pour chaque population.

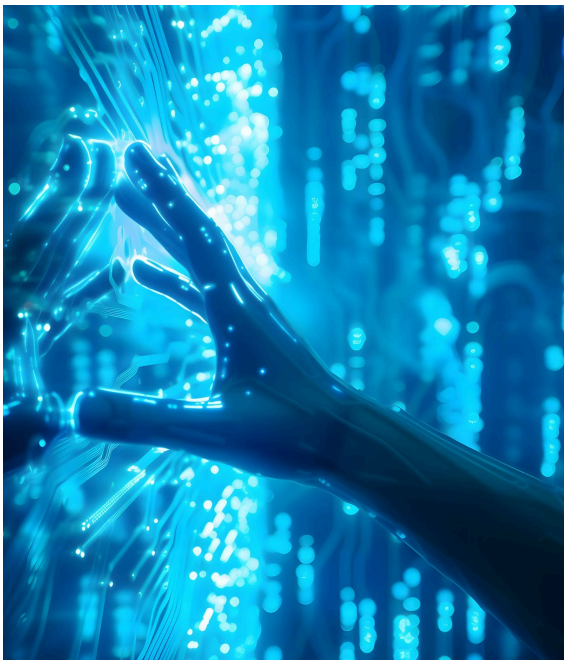
Si le cas d'usage est susceptible d'engendrer un risque de biais ou de discrimination, le GT-IA propose :

- Une ou des populations à protéger, en s'attachant à respecter une cohérence avec les cas d'usage de même nature déjà instruits
- Une métrique d'équité
- Le seuil de tolérance lié à cette mesure.

En cas de dépassement du seuil de tolérance, l'équipe en charge du design du système d'IA instruit une ou plusieurs alternatives visant à remédier ou minimiser ce dépassement.

S'agissant de l'impact environnemental des systèmes d'IA, l'équipe en charge du design adopte une politique d'IA frugale.

Il est demandé de justifier l'utilisation de la technologie employée et d'évaluer le gain marginal visé au regard de l'empreinte environnementale induite. Dans la mesure du possible, l'équipe chargée du design du système d'IA étudie l'alternative de modèles classiques ou de plus faible taille. Dans le cas de systèmes développés en interne, l'équipe documente le cas échéant les optimisations réalisées pour minimiser cette empreinte (quantification, recherche d'hyper paramètres etc.). S'agissant plus spécifiquement des IA génératives, très consommatrices en ressources y compris lors des phases d'utilisation des systèmes, une vigilance renforcée est appliquée. Dans le cas de systèmes SaaS[9], les équipes en charge des systèmes d'IA fournissent leurs meilleurs efforts en termes de veille et de relation fournisseur pour estimer l'impact environnemental des services utilisés.



S'agissant de l'impact social des systèmes d'IA, l'équipe en charge du design documente les articulations homme-machine envisagées. Elle s'astreint à recenser les parties prenantes internes impactées par le système, les éventuels impacts négatifs pouvant découler de leur déploiement, ainsi que leur matérialité et les mesures de remédiation ou minimisation envisagées. Dans les cas d'utilisation d'IA génératives permettant une très grande variété d'interactions homme-machine et multipliant ainsi le risque de mauvais alignement accidentel des objectifs ou d'utilisation détournée par l'humain, l'équipe en charge du design du système documente les risques identifiés, ainsi que les moyens de remédiation ou de minimisation mis en place, notamment s'agissant de la possibilité d'accès, par les utilisateurs, à des informations dangereuses, à la possibilité de comportements dégradants ou abusifs ou encore à la production de résultats toxiques ou biaisés.

[9] Les systèmes SaaS sont des systèmes hébergés par des tiers fournisseurs de service.

Comme mis en avant auparavant, si la plupart des exigences d'un cadre d'IA responsable et de confiance relèvent ou sont minimisées par des mesures qu'elles soient techniques ou opérationnelles, (auditabilité, transparence, interprétabilité, robustesse, performance, etc.), certains enjeux résultent de critères extra techniques pour lesquels un cadre politique mérite d'être mis en place et gouverné. Il est ainsi demandé aux équipes en charge du développement des systèmes d'IA, avec l'appui du GT-IA, de se questionner sur les trois dimensions précédemment citées en mobilisant au besoin les approches conséquentialiste, déontologique et axiologique présentées pour chaque nouveau cas d'usage et, le cas échéant, de solliciter auprès du Comité de Direction l'appui de la Commission Gouvernance du Numérique.

La Commission Gouvernance du Numérique est tenue informée des cas d'usage susceptibles d'être porteurs de risques ou de questions éthiques. En cas de doute nécessitant l'instruction d'une réflexion approfondie, la Commission est saisie. Après instruction, celle-ci se prononce sur les dimensions d'analyse jugées critiques pour le cas d'usage. **Un principe de jurisprudence par typologie de cas d'usage est adopté.** A partir de chaque projet suivi, des recommandations à appliquer aux cas similaires à venir (avec tableau de bord générique) sont élaborées par la Commission Gouvernance du Numérique et partagées aux équipes.

En cas de saisine de la Commission, le Comité de Direction suit l'instruction du dossier, valide les recommandations élaborées et prend en dernière instance la décision du déploiement du système.

Bibliographie

Abrardi, Laura, Carlo Cambini, et Laura Rondi. 2019. « The Economics of Artificial Intelligence: A Survey ».

Aghion, Philippe, Benjamin F. Jones, et Charles I. Jones. 2018. « Artificial Intelligence and Economic Growth ». In *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press, 237-82.

Boodhun, Noorhannah, et Manoj Jayabalan. 2018. « Risk prediction in life insurance industry using supervised learning algorithms ». *Complex & Intelligent Systems* 4(2): 145-54.

Boyd, Ross, et Robert J. Holton. 2018. « Technology, Innovation, Employment and Power: Does Robotics and Artificial Intelligence Really Mean Social Transformation? » *Journal of Sociology* 54(3): 331-45.

Buchanan, Bonnie G. 2019. *Artificial Intelligence in Finance*. The Alan Turing Institute.

Chen, Ying, Catherine Prentice, Scott Weaven, et Aaron Hsiao. 2022. « A systematic literature review of AI in the sharing economy ». *Journal of Global Scholars of Marketing Science* 32(3): 434-51.

Chien, Chen-Fu, Stéphane Dauzère-Pérès, Woonghee Tim Huh, Young Jae Jang, et James R. Morrison. 2020. « Artificial Intelligence in Manufacturing and Logistics Systems: Algorithms, Applications, and Case Studies ». *International Journal of Production Research* 58(9): 2730-31.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, et Aziz Huq. 2017. « Algorithmic Decision Making and the Cost of Fairness ». In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax NS Canada: ACM, 797-806.

Custers, Bart, Toon Calders, Bart Schermer, et Tal Zarsky, éd. 2013. 3 *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-30487-3.

Dubber, Markus D., Frank Pasquale, et Sunit Das. 2020. *The Oxford Handbook of Ethics of AI*. Oxford University Press.

Dupont, Laurent, Olivier Fliche, et Su Yang. 2020. *Gouvernance des algorithmes d'intelligence artificielle dans le secteur financier*. Paris, ACPR.

EIOPA. 2019. *Big Data Analytics in Motor and Health Insurance: A Thematic Review*. Luxembourg: Publications Office of the European Union.

EIOPA. 2019. *Lignes directrices en matière d'éthique pour une IA digne de confiance*. GEHN IA.

EIOPA's Consultative Expert Group on Digital Ethics in insurance. 2021. *Artificial intelligence governance principles: towards ethical and trustworthy artificial intelligence in the European insurance sector*. Luxembourg: Publications Office of the European Union.

Eling, Martin, Davide Nuessle, et Julian Staubli. 2022. « The impact of artificial intelligence along the insurance value chain and on the insurability of risks ». *The Geneva Papers on Risk and Insurance - Issues and Practice* (47): 205-41.

Ethically Aligned Design (v1). 2016. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.

Ethically Aligned Design (v2). 2017. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.

Floridi, Luciano. 2018. « Soft Ethics and the Governance of the Digital ». *Philosophy and Technology* 31(1): 1-8.

Henke, Nicolaus, Jacques Bughin, Michael Chui, James Manyika, Tamim Saleh, Bill Wiseman, et Guru Sethupathy. 2016. *The age of analytics: Competing in a data-driven world*. McKinsey Global Institute.

Ho, Calvin W.L., Joseph Ali, et Karel Caals. 2020. « Ensuring trustworthy use of artificial intelligence and big data analytics in health insurance ». *Bulletin of the World Health Organization* 98(4): 263-69.

Jakšič, Marko, et Matej Marinč. 2019. « Relationship Banking and Information

Technology: The Role of Artificial Intelligence and FinTech ». *Risk Management* 21(1): 1-18.

Jiang, Fei, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, et al. 2017. « Artificial Intelligence in Healthcare: Past, Present and Future ». *Stroke and Vascular Neurology* 2(4).

Keller, Benno. 2020. The Geneva Association *Promoting Responsible Artificial Intelligence in Insurance*.

Kelley, Kevin H., Lisa M. Fontanetta, Mark Heintzman, et Nikki Pereira. 2018. « Artificial Intelligence: Implications for Social Inflation and Insurance ». *Risk Management and Insurance Review* 21(3): 373-87.

Kumar, Ram Shankar Siva, Magnus Nystrom, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, et Sharon Xia. 2020. « Adversarial Machine Learning-Industry Perspectives ». In *2020 IEEE Security and Privacy Workshops (SPW)*, San Francisco, CA, USA: IEEE, 69-75.

Lee, Jay, Hossein Davari, Jaskaran Singh, et Vibhor Pandhare. 2018. « Industrial Artificial Intelligence for Industry 4.0-Based Manufacturing Systems ». *Manufacturing Letters* 18: 20-23.

Li, Bo-hu, Bao-cun Hou, Wen-tao Yu, Xiao-bing Lu, et Chun-wei Yang. 2017. « Applications of Artificial Intelligence in Intelligent Manufacturing: A Review ». *Frontiers of Information Technology & Electronic Engineering* 18(1): 86-96.

Luciano, Elisa, Collegio Carlo Alberto, Matteo Cattaneo, Chief Digital, Innovation Officer, Reale Mutua, et Ron S Kenett. 2023. *Adversarial AI in Insurance : Pervasiveness and Resilience*.

Luciano, Elisa, Behnaz Ameridad, Matteo Cattaneo, et Ron S. Kenett. 2022. « AI and Adversarial AI in insurance: Background, examples, and future implications ». *SSRN Electronic Journal* (667).

Ma, Zixuan, Jiaao He, Jiezhong Qiu, Huanqi Cao, Yuanwei Wang, Zhenbo Sun, Liyan Zheng, et al. 2022. « BaGuaLu: Targeting Brain Scale Pretrained Models with over 37 Million Cores ». In *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, Seoul Republic of Korea: ACM, 192-204.

Makridakis, Spyros. 2017. « The Forthcoming Artificial Intelligence (AI) Revolution: Its Impact on Society and Firms ». *Futures* 90: 46-60.

Manyika, James, Michael Chui, Mehdi Miremadi, Jacques Bughin, Katy George, Paul Willmott, et Martin Dewhurst. 2017. *A Future That Works: Automation, Employment, and Productivity*. McKinsey Global Institute.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, et Aram Galstyan. 2021. « A Survey on Bias and Fairness in Machine Learning ». *ACM Computing Surveys* 54(6).

Mueller, Christoph, et Vitaliy Mezhuyev. 2022. « AI Models and Methods in Automotive Manufacturing: A Systematic Literature Review ». In *Recent Innovations in Artificial Intelligence and Smart Applications*, Studies in Computational Intelligence, éd. Mostafa Al-Emran et Khaled Shaalan. Cham: Springer International Publishing, 1-25.

Mullins, Martin, Christopher P. Holland, et Martin Cunneen. 2021. « Creating ethics guidelines for artificial intelligence and big data analytics customers: The case of the consumer European insurance market ». *Patterns* 2(10): 100362.

Naylor, Michael. 2017. *Insurance Transformed: Technological Disruption*. Cham: Springer International Publishing.

Nilsson, Nils J. 2009. *The Quest for Artificial Intelligence*. Cambridge: Cambridge University Press.

Noordhoek, Dennis. 2023. *Regulation of Artificial Intelligence in Insurance: Balancing Consumer Protection and Innovation*. The Geneva Association.

Patel, Vimla L., Edward H. Shortliffe, Mario Stefanelli, Peter Szolovits, Michael R. Berthold, Riccardo Bellazzi, et Ameen Abu-Hanna. 2009. « The Coming of Age of Artificial Intelligence in Medicine ». *Artificial Intelligence in Medicine* 46(1): 5-17.

Pearce, Adam. 2020. « Measuring Fairness ».

Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, et al. 2016. *« Artificial Intelligence and Life in 2030. » One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016*

Study Panel. Stanford, CA: Stanford University.

Ryll, Lukas, Mary Emma Barton, Bryan Zheng Zhang, R. Jesse McWaters, Emmanuel Schizas, Rui Hao, Keith Bear, et al. 2020. « Transforming Paradigms: A Global AI in Financial Services Survey ». *SSRN Electronic Journal*.

Saleiro, Pedro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, et Rayid Ghani. 2019. « Aequitas: A Bias and Fairness Audit Toolkit ».

Schmidt, Christian. 2018. *Insurance in the Digital Age : A View on Key Implications for the Economy and Society*. The Geneva Association.

Suresh, Harini, et John V. Guttag. 2021. « A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle ». In *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1-9.

Verma, Sahil, et Julia Rubin. 2018. « Fairness Definitions Explained ». In *Proceedings of the International Workshop on Software Fairness*, Gothenburg Sweden: ACM, 1-7.

Aéma Note de position AI Act. 2021. Aéma.

Pour une utilisation responsable et éthique de l'intelligence artificielle dans l'assurance. 2022. France Assureurs.

Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle et modifiant les règlements (CE) n° 300/2008, (UE) n° 167/2013, (UE) n° 168/2013, (UE) 2018/858, (UE) 2018/1139 et (UE) 2019/2144 et les directives 2014/90/UE, (UE) 2016/797 et (UE) 2020/1828 (règlement sur l'intelligence artificielle).

Contacts presse

Marina DUCROS - 06 13 55 57 98 - mducros@macif.fr

Joanne BENHAIM - 06 62 65 11 66 - jbenhaim@macif.fr

MACIF - MUTUELLE ASSURANCE DES COMMERÇANTS ET INDUSTRIELS DE FRANCE ET DES CADRES ET SALARIÉS DE L'INDUSTRIE ET DU COMMERCE. Société d'assurance mutuelle à cotisations variables. Entreprise régie par le Code des assurances. Siège social : 1 rue Jacques Vandier 79000 Niort.